ATHLATES User Manual 1.0

Xiao Yang[1] and Chang Liu[2]

[1]Genome Sequencing and Analysis Program, Broad Institute of MIT and Harvard
[2]Department of Pathology and Immunology, Washington University School of Medicine, Washington University St. Louis

July 31, 2014

Contents

1 General Description

As we expect that high coverage exome-seq will be routinely carried out for individual patient in clinical settings, there would be no additional cost to infer HLA types for the individual conditional on that both alleles for each target HLA-gene were sufficiently captured. This assumption is expected to be valid as the quality of probes improve. ATHLATES is created to fill in this needs. It is mainly relying on assembly, which attempts to create highly reliable haplotypes of exons that could provide better phasing information compared with read alignments. ATHLATES generates intuitive reports that can be directly used by clinicians.

2 Prepare input for ATHLATES

1. Collect all HLA cDNA and genomic sequences from the HLA/IMGT database to generate a multi-fasta file [hlall.fa]. For example, in the ATHLATES release folder, this file is named [hla.clean.fasta] in the db/ref/ folder.

2. Align input paired Fastq reads to the [hlall.fa], *e.g.*,

   $ mpirun -machinefile machinefile -perhost 2 -np 9 novoalignMPI -d [ref.nix] -f [fwd.fq] [rv.fq] -t 30 -o SAM -r all -l 80 -e 100 -i PE 200 140 | samtools view -bS -h -F 4 - > [output.bam]

   NOTE:
   .nix – novoalign index file "$ novoindex ref.nix hlall.fa"
   -F 4 – skip unmapped reads when converting sam to bam.
   -e 100 -r all – if a read aligns to the genome up to 100 places equally well, all of those alignments will be recorded (Novoalign will stop trying to align the read after 100 alignments are found). When -r random, a single alignment will be picked.
   -i PE 200 140 – fragment length 200, std 140. This parameter needs to be adjusted according to your library information.

   ALTERNATIVELY, open source aligner MOSAIK (2.1.73) can be used here and we found that by allowing no mismatches (for novoalign, this means reducing -t e.g.: -t 10) while tolerating soft-clipping yield better assembly results in ATHLATES. We used the following parameters for MOSAIK: -minp 0.4 -p 10 -mms -3 -ms 1 -hgop 4 -gop 5 -gep 2 -m all -bw 29 -a all -act 20 -mm 0

3. Sort.
   $ samtools sort [output.bam] [output.sort]

4. Extract an HLA-gene specific reads and reads not belonging to this gene. The resulting BAM file is sorted primarily by read name, and secondarily by reference name. This is applied to both reads not belonging to this gene.

   $ samtools view -b -L A_nucgen.bed [output.sort.bam] > [A.bam]
   $ samtools view -h -o [A.sam] [A.bam]

```
IMGT/HLA Release: 3.9.0
Sequences Aligned: 2012 July 12
Steven GE Marsh, Anthony Nolan Research Institute.
Please see http://hla.alleles.org/terms.html for terms of use.

 cDNA              1
 AA codon          -24
                   |
 C*01:02:01        ATG CGG GTC ATG GCG CCC CGA ACC CTC ATC CTG CTG CTC TCG GGA GCC CTG GCC CTG ACC GAG ACC TGG GCC T|GC
 C*01:02:02        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:03        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:04        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:05        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:06        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:07        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:08        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:09        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:10        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:11        --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:02:12        --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:02:13        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:14        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:02:15        --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:02:16        *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:03           --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:04           --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:05           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:06           --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:07           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:08           --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- --- -|--
 C*01:09           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:10           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:11           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
 C*01:12           *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *** *|--
```

Figure 1: An example MSA file of HLA-C in IMGT/HLA database.

$ sort -k 1,1 -k 3,3 [A.sam] > [A.sort.sam]
$ samtools view -bS [A.sort.sam] > [A.sort.bam]

$ samtools view -b -L non_A.bed [output.sort.bam] > [non-A.bam]
$ samtools view -h -o [non-A.sam] [non-A.bam]
$ sort -k 1,1 -k 3,3 [non-A.sam] > [non-A.sort.sam]
$ samtools view -bS [non-A.sort.sam] > [non-A.sort.bam]

NOTE:
A_nucgen.bed and non_A.bed are pre-curated bed file and they should be updated whenever [hlall.fa] is updated. The former and the latter keep records of HLA-A and non-HLA-A gDNA and cDNA sequences in [hlall.fa], respectively.

5. The multiple sequence alignment file should be downloaded from IMGT/HLA database. An example of such a file is shown in Fig. 1 for HLA-C, where the MSA file is named as [C_nuc.txt]

3 Quick Start

3.1 Pre-requisite

1. Installation of BamTools (version e235c55 or later) (instructions can be found here `https://github.com/pezmaster31/bamtools/wiki`).
2. Installation of Perl (recent versions are recommended)
3. g++ compiler (recent versions are recommended)

3.2 Procedure

1. Download the ATHLATES package, decompress, and "cd" into the "Athlates-version" folder.
2. Switch to bash environment.

   $ bash

3. Export BamTool library path. Assuming you successfully installed BamTool-e235c55 in directory [path], then you should be able to find the library in directory "[path]/lib"

   $ export `LD_LIBRARY_PATH=$LD_LIBRARY_PATH:/[path]/lib`

4. Compile and Run ATHLATES
   (a) Edit file "Athlates/src/makefile" –
      - set MYPATH to be [path], *e.g.* MYPATH=/MyLibrary/bamtools-e235c55/
      - set COMPILER to be the path of the g++ compiler you are using (you could use command "$ which g++" to find out this information).

   (b) Compile in "src" directory

      $ cd src
      $ make
      $ cd ../

      Note:
      - The executive file can be found in the "bin" folder.

   (c) Run ATHLATES

      $ ./bin/typing

      This will show you options to set parameters for the program.

      $ ./bin/typing -bam [A.sort.bam] -exlbam [non-A.sort.bam] -msa [A_nuc.txt] -o [Out-Prefix]

      Parameters:

- -bam – input sorted (by -n) BAM file for the target HLA gene.
- -exlbam – input BAM file for reads aligned to non-target HLA genes.
- -msa – the MSA file of the target HLA gene.
- -o – output prefix name.

Please replace the values between [ ]

5. Output

   (a) {OutPrefix}.typing.txt – typing result. An example is given in Fig. 2. Typically, user would only be interested in this file.
   (b) {OutPrefix}.contig.fa – assembled contig file in fasta format.
   (c) {OutPrefix}.contig.detail.txt – same as the previous one but with additional information of coverage.
   (d) {OutPrefix}.unpair.fa – paired end reads that were merged.
   (e) {OutPrefix}.pair.fa – paired end reads that failed to be merged.
   (f) {OutPrefix}.raw.fa – all the unique reads included in the input [.sort.bam] file.

6. Interpretation of typing results. As shown in Fig. 2. There are three parts in the output. In the first part, a list of potential alleles that are considered to be supported by the data is shown. HD denotes the Hamming distance between this allele to the contigs. Note that for each allele, either due to incomplete coverage of certain exon or the exon has large distance compared to to the assembled contigs (due to insertion or deletion mutation for instance), the corresponding exon is excluded from HD calculation. For example, for allele DQB1*03:03:02:03, exon 6 with 14bp long is excluded. In this case, as the exon is very small, most likely, it has not been captured. The second part provides a subset of candidate alleles chosen from the first part to be considered during allelic pair inference. And the last part provide all possible typing results. When both alleles of a target HLA gene were not sufficiently captured (by default $\leq$ 85% of the full cDNA) in exome-seq, Athlates will not report any typing results. However, in the case when only one allele is captured out of the heterozygous alleles, Athlates will report the typing results to be a homozygous. This might also occur with the conventional typing (Sanger sequencing of amplicons from HLA loci) when one haplotype fails to be amplified. Therefore, homozygous alleles are commonly confirmed by another independent method; for example, the reverse SSO DNA typing is routinely used in this situation in our laboratory. Finally, the reliably constructed allelic pair will have score value of 0. Any non-zero score value is likely reflecting incomplete capturing of full length HLA gene. In these cases, further investigation can be carried out using the typing report as well as other files created by Athlates (see previous step).

4 Demo - a walk through

We have prepared a "demo" folder in the release to provide a walk through for applying Athlates to exome-seq data of individual HG01756 from 1000 genome project.

1. Preparing input.

Name	HD	Aln_len	cDNA_len	Similarity	Avg_cov	Missing Exons (ID, len) ; mismatches [ID, pos]
DQB1*03:03:02:03	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*03:03:02:02	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*03:03:02:01	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*02:01:01	0	772	786	1	144.622	(5, 0) (6, 14)
DQB1*02:02	1	772	786	0.998705	144.622	(5, 0) (6, 14) [3, 121]
DQB1*03:02:01	1	772	786	0.998705	134.253	(5, 0) (6, 14) [2, 157]
DQB1*03:31	1	618	618	0.998382	147.324	(1, 0) (5, 0) (6, 0) [3, 63]
DQB1*03:43	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [3, 141]
DQB1*03:41	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [2, 120]
DQB1*03:39	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [3, 98]
DQB1*03:38	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [2, 75]
DQB1*03:30	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [2, 13]
DQB1*03:03:04	1	552	552	0.998188	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [3, 224]
DQB1*02:01:04	1	552	552	0.998188	173.219	(1, 0) (4, 0) (5, 0) (6, 0) [2, 194]
DQB1*02:01:05	1	552	552	0.998188	173.219	(1, 0) (4, 0) (5, 0) (6, 0) [2, 206]
DQB1*02:07	1	552	552	0.998188	173.219	(1, 0) (4, 0) (5, 0) (6, 0) [2, 229]
DQB1*02:04	1	552	552	0.998188	173.219	(1, 0) (4, 0) (5, 0) (6, 0) [3, 123]
DQB1*03:33	1	522	522	0.998084	149.414	(1, 0) (4, 0) (5, 0) (6, 0) [3, 156]
DQB1*03:34	1	522	522	0.998084	149.414	(1, 0) (4, 0) (5, 0) (6, 0) [2, 45]
DQB1*03:32	2	552	552	0.996377	154.324	(1, 0) (4, 0) (5, 0) (6, 0) [2, 157] [3, 138]
DQB1*02:06	2	552	552	0.996377	173.219	(1, 0) (4, 0) (5, 0) (6, 0) [3, 121] [3, 225]

------------ Candidate Allelic Pairs -------------

Name	HD	Aln_len	cDNA_len	Similarity	Avg_cov	Missing Exons (ID, len) ; mismatches [ID, pos]
DQB1*03:03:02:03	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*03:03:02:02	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*03:03:02:01	0	772	786	1	134.253	(5, 0) (6, 14)
DQB1*02:01:01	0	772	786	1	144.622	(5, 0) (6, 14)

------------ Inferred Allelic Pairs -------------

DQB1*03:03:02:03	DQB1*02:01:01	0
DQB1*03:03:02:02	DQB1*02:01:01	0
DQB1*03:03:02:01	DQB1*02:01:01	0

HD: Hamming distance.
Aln_len: alignment length, number of cDNA bases supported by contigs.
cDNA_len: total length of cDNA of an allele.
Avg_cov: average coverage for an allele.
Missing Exons (ID, len): ID, the identity of a exon not considered for calculation of Hamming distance; len, length of the indicated exon as documented in the IMGT/HLA database.
mismatches[ID, pos]: the position (pos) of a mismatch in the exon of indicated identity (ID) when compared to its best hit in contigs.

Figure 2: An example ALTHLATES output.

- Collecting references, which could be found at: "db/ref/hlall.fa"; [.bed] files for each target and off-target HLA-genes were then created and can be found in folder "db/bed/"
- The raw fastq paired end read files were downloaded from 1000 genome project ftp site and aligned using novoalignMPI as shown in section 2. The resulting BAM file can be found at "demo/HG01756/HG01756hlall.bam"
- Following section 2 step 3 and 4, we obtain corresponding target and non-target BAM files for each HLA gene. These files can be found in folder "demo/HG01756"
- The MSA file for each target HLA-gene were obtained from the IMGT/HLA database and can be found in folder "db/msa"

2. Assuming Athlates is compiled properly, run Athlates.

   $ ./bin/typing -bam demo/HG01756/HG01756_a.sort.bam -exlbam demo/HG01756/HG01756_na.sort.bam -msa db/msa/A_nuc.txt -o demo/output/HG01756_a > demo/output/HG01756_a.log.txt

3. The output can be found in the folder "demo/output/"

5 License

Please refer to license folder.

6 Citing ATHLATES

C. Liu, X. Yang, B. Duffy, T. Mohanakumar, R.D. Mitra, M.C. Zody, J.D. Pfeifer (2012) "ATHLATES: accurate typing of human leukocyte antigen through exome sequencing", Nucl. Acids Res. (2013) [doi: 10.1093/nar/gkt481]

7 Contact

If you have any question, please email Xiao Yang (xiaoyang@broadinstitute.org).